Institut royal des Sciences naturelles de Belgique

BULLETIN

Tome XXXI, n° 56
Bruxelles, novembre 1955.

Koninklijk Belgisch Instituut voor Natuurwetenschappen

MEDEDELINGEN

Deel XXXI, nr 56
Brussel, November 1955.

MESURE DE DIVERGENCE Δ^2 ENTRE UN SUJET DÉTERMINÉ ET UNE POPULATION MULTIVARIÉE NORMALE. SA DISTRIBUTION D'ÉCHANTILLONNAGE,

par Elisabeth DEFRISE-GUSSENHOVEN (Bruxelles).

INTRODUCTION.

Grâce aux progrès des statistiques mathématiques, les méthodes biométriques appliquées à l'anthropologie permettent actuellement de faire la synthèse entre, d'une part, les nombreuses données numériques, les longues listes de mensurations dont l'auteur n'arrivait pas toujours à dégager des lois générales et, d'autre part, les descriptions qualitatives, très fines, très poussées, mais qui restaient néanmoins sujettes à caution et ne situaient pas nettement l'objet étudié dans sa lignée évolutive ou n'en précisaient pas la place dans l'humanité actuelle.

Aussi, croyons-nous que la biométrie contribuera à faire sortir l'anthropologie de l'ornière où celle-ci est plus ou moins engagée, en garantissant une grande précision dans la description des documents et en condensant les mensurations en graphiques dont les lois se dégagent aisément. En même temps, l'objectivité de telles méthodes assurera une large diffusion à chaque découverte en fournissant une base de discussion solide et concrète aux conclusions de l'auteur.

C'est dans cet esprit que le professeur F. TWIESSELMANN (1) a abordé l'étude du fémur de Fond-de-Forêt; ainsi, pour rendre plus éloquente la description du fossile, il a comparé celui-ci avec les autres fémurs néanderthaliens et avec une collection de fémurs actuels.

Un problème s'est alors posé, qui, à ma connaissance, n'est pas résolu en biométrie : évaluer l'écart entre un spécimen isolé et une population homogène, connue par un échantillon.

Il n'y a pas lieu ici d'appliquer un test d'appartenance puisque le fossile n'appartient pas à une population récente (2).

Ce qu'il faut est une mesure de divergence qui permette d'exprimer par un nombre l'éloignement du fémur de Fond-de-Forêt de celui de l'Homme moderne. Une telle notion de « distance » doit servir en outre à comparer les écarts d'autres fémurs fossiles et à réunir toutes les données relatives à cette question en un graphique significatif.

P. C. MAHALANOBIS (3) a traité une question analogue quand il a introduit la « distance » entre les moyennes de deux populations dans le but d'améliorer le coefficient de ressemblance raciale. Il a créé une quantité $\triangle^2$, appelée actuellement distance généralisée de MAHALANOBIS, qui dépend des moyennes, des variances et des covariances des populations supposées normales.

Nous inspirant de cet exemple, nous avons défini une mesure de divergence Λ^2 qui exprime le carré de la « distance » entre un point D (représentant le fossile) et un point M qui est l'image des moyennes de la population de référence, supposée normale.

Λ^2 est défini par une expression mathématique assez complexe. Aussi, pour rendre cette notion de distance plus accessible, nous appelons « taux d'éloignement » de D le pourcentage des sujets de la population de référence qui sont plus près du point moyen M que le spécimen D à l'étude; autrement dit, le taux

(1) TWIESSELMANN, F., 1954, et un mémoire à paraître dans les publications de l'Institut royal des Sciences naturelles de Belgique.

(2) Plusieurs auteurs ont attiré l'attention sur la distinction essentielle entre les notions de test et mesure de divergence; voir notamment MAHALANOBIS, P. C., 1930; FISHER, R. A., 1936; SELTZER, C. C., 1937.

(3) MAHALANOBIS, P. C., 1936. Voir aussi RAO, R. C., 1952, p. 355, où l'auteur montre les avantages de la distance généralisée sur l'ancien coefficient de ressemblance raciale de K. Pearson.

d'éloignement est le pourcentage de sujets du groupe dont le Λ^2 est inférieur à celui du sujet isolé D. Il est nul si le fossile D se trouve au centre de la population de référence.

Dans deux notes précédentes (4), nous avons déjà défini le taux d'éloignement ainsi que le Λ^2 et signalé diverses applications en paléontologie humaine, en anthropologie physique et en systématique animale. La première note avait un but essentiellement pratique, de sorte que nous nous sommes particulièrement arrêté au cas de deux variables qui se prête à la construction de graphiques, où la population de référence est entourée d'ellipses équiprobables. Cette représentation est très commode; plusieurs naturalistes de l'Institut royal des Sciences naturelles de Belgique l'emploient couramment.

Dans la note actuelle, nous traitons la question d'un point de vue plus général. Nous définissons la mesure de divergence et le taux d'éloignement pour p variables normales; nous calculons la répartition d'échantillonnage de L^2 et les moments de cette distribution afin de pouvoir juger la validité du Λ^2 et du taux d'éloignement dans le cas où la population de référence ne serait connue que par un nombre réduit de spécimens. L^2 est la valeur correspondant à Λ^2, mais calculée pour l'échantillon; c'est à partir de L^2 que nous estimerons Λ^2.

1. — Définition de Λ^2, mesure de divergence entre un point fixe D et le point moyen (centre) d'une population.

Soit une population multivariée normale non singulière relative à p caractères. Désignons par m_i et α_{ij} $(i,j=1,...,p)$ les moyennes et les covariances $(\alpha_{ij}=\rho_{ij}\sigma_{ij})$. Soient $\bar{x}_i$ et $a_{ij}=r_{ij}s_i s_j$ les moyennes et les covariances d'un échantillon d'effectif n tiré de cette population et d_i les coordonnées d'un point fixe D dont nous voulons définir la « distance » au centre M de la population.

Nous appelons

$$\Lambda^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij}(m_i - d_i)(m_j - d_j)$$

(4) Defrise-Gussenhoven, E., 1955.

la mesure de divergence (ou le carré de la « distance ») entre le point D et le centre M de la population, α^{ij} étant le mineur normé de l'élément α_{ij} dans la matrice des covariances $A = ||\alpha_{ij}||$.

Quand le point D varie, la quantité Λ^2 est distribuée comme χ^2 avec p degrés de liberté. Les tables usuelles de χ^2 indiquent donc immédiatement le pourcentage de sujets de la population qui ont un Λ^2 inférieur ou égal à celui du point D.

Ce pourcentage est, par définition, le taux d'éloignement du point D par rapport à la population.

2. — Justification de la définition de Λ^2.

Pour représenter une « distance », il faut qu'une fonction satisfasse à certaines conditions, énoncées par P. C. Mahalanobis (5) et C. R. Rao (6).

Nous allons montrer que Λ^2 répond à ces exigences.

1° Λ^2 est un scalaire non négatif. Il est nul lorsque D est confondu avec le centre M de la population et augmente indéfiniment à mesure que D s'éloigne de M dans une direction donnée.

2° Λ^2 est invariant pour les transformations linéaires qui affectent les variables. En particulier, Λ^2 a donc la même valeur quelles que soient les unités de mesure adoptées pour représenter les mensurations.

Démonstration. Utilisons la notation matricielle (7).

La distribution des p variables de la population est

$$\text{const.} \times \mathbf{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})' A^{-1} (\mathbf{x}-\mathbf{m})} \prod_{i=1}^{p} dx_i \quad \text{avec} \quad A^{-1} = ||\alpha^{ij}||$$

les coordonnées du vecteur $\mathbf{x}$ étant $x_1, \ldots, x_p$ et celles de $\mathbf{m}$ $m_1, \ldots, m_p$.

(5) Mahalanobis, P. C., 1930, p. 548.
(6) Rao, C. R., 1952, p. 351.
(7) Voir par exemple Cramèr, H., 1946, p. 103. Dans cette notation, $\mathbf{x}$ est un vecteur-colonne, $\mathbf{x}'$ un vecteur-ligne.

Une transformation linéaire, de matrice C, conduit à une nouvelle distribution,

$$\text{const.} \times e^{-\frac{1}{2}(\mathbf{y}-\mathbf{m}^*)'B^{-1}(\mathbf{y}-\mathbf{m}^*)} \prod_{i=1}^{p} dy_i$$

avec $\mathbf{x}=C\mathbf{y}$, $\mathbf{m}=C\mathbf{m}^*$ et $B^{-1}=C'A^{-1}C$.

La matrice B est celle des covariances des nouvelles variables.

On a $\Delta^2=(\mathbf{m}-\mathbf{d})'A^{-1}(\mathbf{m}-\mathbf{d})=(\mathbf{m}^*-\mathbf{d}^*)'C'A^{-1}C(\mathbf{m}^*-\mathbf{d}^*)$ $(\mathbf{d}=C\mathbf{d}^*)$

Le dernier membre de ces égalités peut s'écrire aussi

$$(\mathbf{m}^*-\mathbf{d}^*)'B^{-1}(\mathbf{m}^*-\mathbf{d}^*)$$

qui est le Δ^2 calculé pour les nouvelles variables $y_1,\dots,y_p$.

Il résulte de cette propriété d'invariance de Δ^2 que l'on peut supposer, sans nuire à la généralité, $\alpha_{ij}=0(i\neq j)$ et $\alpha_{ii}=1$. Il suffit de choisir une transformation linéaire qui réduise la forme quadratique définie positive $(\mathbf{x}-\mathbf{m})'A^{-1}(\mathbf{x}-\mathbf{m})$ à une somme de carrés. Alors Δ^2 s'écrira

$$\Delta^2=(m^*_1-d^*_1)^2+\dots(m^*_p-d_p^*)^2.$$

3° Δ^2 ne diminue pas quand on augmente le nombre de variables. En effet, quel que soit le nombre de variables, la forme quadratique $(\mathbf{x}-\mathbf{m})'A^{-1}(\mathbf{x}-\mathbf{m})$ est toujours définie positive; il est possible de trouver une suite de transformations linéaires qui réduisent Δ^2 à une somme de carrés, de telle sorte que l'addition de chaque nouvelle variable entraîne l'addition d'un nouveau terme $(m^*_i-d^*_i)^2$ qui n'est jamais négatif.

4° Ajoutons une dernière condition qui nous semble indispensable à toute mesure de divergence et à laquelle satisfait Δ^2 : la valeur L^2 de Δ^2 calculée à partir d'un échantillon a une distribution d'échantillonnage exacte, ce qui permet d'estimer convenablement Δ^2 à partir de L^2 et d'évaluer les erreurs d'échantillonnage, même lorsque l'effectif est faible.

Remarque 1. La définition de « distance » que nous venons de donner peut s'étendre à tous les couples de points de l'espace. On appellera carré de la « distance » entre deux points F et G, le scalaire

$$(\mathbf{f}-\mathbf{g})'A^{-1}(\mathbf{f}-\mathbf{g})$$

f_i et g_i étant respectivement les coordonnées de F et G.

L'invariance de ce scalaire pour les transformations linéaires (qui se démontre comme celle de Λ^2) entraîne la propriété suivante :

« Distance » de FG + « distance » de GH $\geqslant$ « distance » de FH, F. G et H étant trois points quelconques de l'espace.

Pour le vérifier, il suffit de supposer $\alpha_{ij}=0 (i \neq j)$ et $\alpha_{ii}=1$, car dans ce cas les « distances » définies se confondent avec les distances ordinaires de l'espace euclidien.

Cette propriété sera utile quand nous voudrons comparer deux fossiles, non seulement au point moyen de la population de référence, mais encore entre eux.

Remarque 2. On peut encore donner une autre interprétation géométrique de Λ^2. On considère un système de référence constitué par p vecteurs dont les produits scalaires sont α_{ij}. Dans cet espace, $|\Lambda|$ représente la distance géométrique ordinaire du point D au centre M de coordonnées $m_1,..,m_p$ et l'équation

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij}(x_i-m_i)(x_j-m_j)=c^2 \text{ (constante)}$$

représente une hypersphère de centre M et de rayon c. L'invariance de Λ^2 se démontre aisément à l'aide de cette représentation.

Remarque 3. Par sa nature même, la mesure de divergence conduit à un « taux d'éloignement » qui peut fournir un test d'appartenance du point D à la population. Le taux d'éloignement est une fonction de Λ^2, mais comme il ne varie que de 0 à 1, les points éloignés du centre de la population auront tous un taux d'éloignement pratiquement égal à 1. D'autre part, si l'addition d'une nouvelle variable ne diminue jamais Λ^2, elle peut avoir pour effet de diminuer le taux d'éloignement. Ces raisons nous font considérer Λ^2 comme une bonne mesure de divergence, tandis que le taux d'éloignement a plutôt une valeur pratique, notamment lorsqu'il s'agit d'examiner des points assez proches du centre de la population.

3. — UTILISATION DE Λ^2 DANS LFS PROBLÈMES DE DISCRIMINATION.

Dans les problèmes de discrimination, on doit choisir entre deux populations pour y classer un sujet. R. A. FISHER a donné une solution dans le cas où les deux populations sont multivariées normales, avec des moyennes différentes, mais avec une même matrice de variances et covariances.

La frontière entre les deux populations est alors donnée par une fonction discriminatoire linéaire des variables, telle que le nombre de sujets mal classés est minimum (8). Dans une note précédente (9), nous avons montré que dans le cas de deux variables, la fonction discriminatoire représente la droite joignant les points d'intersection de deux ellipses équiprobables qui contiennent respectivement un même pourcentage de sujets de chacune des deux populations.

Grâce au Λ^2, il est possible d'éliminer l'hypothèse de l'égalité des variances et covariances dans les deux populations.

En effet, soient deux populations multivariées normales P_1 et P_2 dont les distributions sont respectivement

$$C_1 \times e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)' A^{-1}_1 (\mathbf{x}-\mathbf{m}_1)} \Pi dx_i$$

et

$$C_2 \times e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_2)' A^{-1}_2 (\mathbf{x}-\mathbf{m}_2)} \Pi dx_i$$

et soient $d_1,..,d_p$ les coordonnées d'un sujet D à classer dans l'une des deux populations.

On calcule le carré de la « distance » Λ^2 de D par rapport à P_1 et P_2

$$\Lambda^2_1 = (\mathbf{m}_1 - \mathbf{d})' A^{-1}_1 (\mathbf{m}_1 - \mathbf{d}) \quad \text{et} \quad \Lambda^2_2 = (\mathbf{m}_2 - \mathbf{d})' A^{-1}_2 (\mathbf{m}_2 - \mathbf{d}).$$

On classera le sujet dans P_1 si $\Lambda^2_1 < \Lambda^2_2$ et dans P_2 si $\Lambda^2_2 < \Lambda^2_1$.

Le lieu des points pour lesquels $\Lambda^2_1 = \Lambda^2_2$ a pour équation

$$(*) \qquad (\mathbf{x}-\mathbf{m}_1)' A^{-1}_1 (\mathbf{x}-\mathbf{m}_1) = (\mathbf{x}-\mathbf{m}_2)' A^{-1}_2 (\mathbf{x}-\mathbf{m}_2).$$

(8) FISHER, R. A., 1936 et 1937.
(9) DEFRISE-GUSSENHOVEN, E., 1952, p. 27.

C'est une hyperquadrique lorsque $A_1 \neq A_2$; on retrouve la fonction discriminatoire de R. A. FISHER lorsque $A_1 = A_2$, car alors les termes du 2me degré de (*) se détruisent.

De même, si l'on a le choix pour classer le sujet D entre k populations multivariées normales $P_1, \ldots, P_k$, à moyennes et covariances distinctes, on calculera les carrés des « distances » $\Lambda^2_1, \ldots, \Lambda^2_k$ de D à chacune des populations. On classera le sujet D dans la population P_i correspondant à la valeur minimum Λ^2_i.

Il resterait évidemment à déterminer le nombre de sujets mal classés dans le cas de deux populations multivariées à covariances inégales. Ce problème requiert l'intégration de la densité de la population P_1 d'un même côté de l'hyperquadrique d'équation (*).

4. — RECHERCHE DE LA DISTRIBUTION ÉCHANTILLONNÉE DE L^2, VALEUR DE Λ^2 CALCULÉE A PARTIR D'UN ÉCHANTILLON.

Calculée à partir d'un échantillon d'effectif n, la mesure de divergence Λ^2 vaut

$$L^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} a^{ij} (\overline{x}_i - d_i)(\overline{x}_j - d_j)$$

fonction des moyennes et des covariances de l'échantillon. Nous allons établir la répartition d'échantillonnage de L^2 quand l'échantillon tiré de la population varie, le point D étant fixe.

Faisons appel à un théorème démontré par C. R. RAO (10).

a) Théorème de RAO : Considérons un échantillon d'effectif n tiré d'une population normale à p variables dont la répartition est

$$\text{const.} \times e^{-\frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \beta^{ij} (y_i - \mu_i)(y_j - \mu_j)} \Pi dy_i$$

(10) RAO, C. R., 1952, p. 72.

Posons

$$S_p = \frac{|S_{ij}|}{|S_{ij} + n\overline{y}_i\overline{y}_j|},$$

où

$$S_{ij} = \sum_{k=1}^{n} (y_{ik} - \overline{y}_i)(y_{jk} - \overline{y}_j)$$

et où les $\overline{y}_i$ sont les moyennes de y_i pour l'échantillon; la répartition d'échantillonnage de S_p est alors

$$(**) \quad \text{const.} \times S_p^{(n-p-2)/2}(1-S_p)^{(p-2)/2}{}_1F_1\left(\frac{n}{2}, \frac{p}{2}, \frac{(1-S_p)n\mu^2}{2}\right)dS_p$$

où

$$\mu^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \beta^{ij}\mu_i\mu_j$$

et où ${}_1F_1$ est la fonction hypergéométrique définie par

$${}_1F_1(\alpha,\rho,z) = 1 + \frac{\alpha}{1!\rho}z + \frac{\alpha(\alpha+1)}{2!\rho(\rho+1)}z^2 + \frac{\alpha(\alpha+1)(\alpha+2)}{3!\rho(\rho+1)(\rho+2)}z^3 + \ldots$$

La démonstration de cette propriété est obtenue par C. R. Rao grâce à une suite de théorèmes très bien agencés qui servent aussi dans d'autres questions, notamment celle de la distribution du T^2 de Hotelling, du D^2 de Mahalanobis, etc... (11).

b) Application du théorème de C. R. Rao à la répartition de L^2.

Supposons que le vecteur y de la population envisagée par Rao soit égal à

$$\mathbf{y} = \mathbf{x} - \mathbf{d}$$

(11) Fisher, R. A. 1915 et 1928.
Hotelling, H. 1931.
Bose, R. C. et Roy, S. N. 1938.
Roy, S. N. 1939.

où les composantes de $\mathbf{x}$ et $\mathbf{d}$ sont respectivement les variables x_i et les constantes d_i définies au paragraphe 1 des définitions.

Posons en outre

$$||\beta_{ij}||=||\alpha_{ij}|| \quad \text{et} \quad \mu_i=m_i-d_i \qquad (i,j=1,...,p)$$

les m_i étant les moyennes et les α_{ij} étant les covariances des variables x_i.

Alors $S_{ij}=na_{ij}$, où les a_{ij} désignent comme au paragraphe 1 les covariances des x_i dans l'échantillon. Dès lors, S_p peut s'écrire

$$S_p=\frac{|na_{ij}|}{|na_{ij}+n(\overline{x}_i-d_i)(\overline{x}_j-d_j)|}=\frac{1}{1+L^2}$$

car on montre aisément que

$$L^2=\begin{vmatrix} o & -(\overline{x}_1-d_1) & . . & -(\overline{x}_p-d_p) \\ (\overline{x}_1-d_1) & a_{11} & . . & a_{p1} \\ . & . & . . & . \\ (\overline{x}_p-d_p) & a_{p1} & . . & a_{pp} \end{vmatrix} : |a_{ij}|$$

c) Répartition de L^2.

Pour avoir la répartition de L^2, il suffit de faire dans (**)

$$S_p=\frac{1}{1+L^2} \quad \text{et} \quad dS_p=-(1+L^2)^{-2}dL^2$$

La constante se calcule aisément par intégration. Quand S_p varie de 0 à 1, L^2 varie de ∞ à 0. La répartition d'échantillonnage de L^2 prend la forme

$$\frac{\Gamma(n/2)e^{-n\Lambda^2/2}}{\Gamma(p/2)\Gamma({}^{(n-p)}/2)} \frac{(L^2)^{(p-2)/2}}{(1+L^2)^{n/2}} {}_1F_1\left(\frac{n}{2},\frac{p}{2},\frac{n\Lambda^2L^2}{2(1+L^2)}\right) dL^2$$

L^2 variant de 0 à ∞ et $\Lambda^2=(\mathbf{m}-\mathbf{d})'\mathrm{A}^{-1}(\mathbf{m}-\mathbf{d})$ étant le « carré de la distance » du point D au centre de la population.

Remarques.

1. Cette distribution ne dépend que de l'effectif n de l'échantillon, du nombre de variables p et de Λ^2.

2. Quand $\Lambda^2=0$, c'est-à-dire quand le point D se trouve au centre de la population, la distribution se réduit à

$$\frac{\Gamma(n/2)}{\Gamma(p/2)\Gamma((n-p)/2)}\frac{(L^2)^{(p-2)/2}}{(1+L^2)^{n/2}}dL^2$$

et permet de tester l'hypothèse que le centre de la population dont on a tiré l'échantillon se confond avec le point D. Il suffit de poser

$$L^2=\frac{T^2}{n-1}$$

pour retrouver la distribution du T^2 de HOTELLING.

3. Quand $p=1$, on n'a qu'une seule variable x, ayant une distribution N (m,σ); on a alors

$$\Lambda^2=\frac{(m-d)^2}{\sigma^2} \quad \text{et} \quad L^2=\frac{(\bar{x}-d)^2}{s^2};$$

$\bar{x}$ et s sont respectivement la moyenne et la déviation standard d'un échantillon et d la coordonnée du point D dont on cherche la mesure de divergence à la moyenne m. Dans ce cas, la répartition de L^2 devient celle du t^2 non central de STUDENT, à condition de faire

$$L^2=\frac{t^2}{n-1}.$$

5. — Moments de la répartition de L^2 et estimation de Λ^2.

Les premier et deuxième moments se calculent par intégration.

$$m(L^2) = \frac{p+n\Lambda^2}{n-p-2};$$

$$m_2(L^2) = \frac{p^2+2p+2(p+2)n\Lambda^2+n^2\Lambda^4}{(n-p-2)(n-p-4)} = \frac{(p+2n\Lambda^2)(p+2)+n^2\Lambda^4}{(n-p-2)(n-p-4)}$$

La variance de L^2 est

$$\sigma^2(L^2) = \frac{2n^2\Lambda^4+2(n-2)(p+2n\Lambda^2)}{(n-p-2)^2(n-p-4)}$$

Un estimateur sans biais et consistant de Λ^2 est

$$\Lambda^2_e = \frac{n-p-2}{n} L^2 - \frac{p}{n}$$

L^2 est un estimateur biaisé mais consistant de Λ^2.

Il faut choisir comme estimateur de Λ^2 :

$$\Lambda^2_e = \frac{(n-p-2)L^2}{n} - \frac{p}{n}$$

L'étude et la tabulation de la distribution de L^2, qui dépend de trois paramètres n, p et Λ^2, restent à faire et permettraient de déterminer dans chaque cas les intervalles de confiance de l'estimateur Λ^2_e (12).

(12) Nous avons déjà utilisé la mesure de divergence Λ^2 dans une question de paléontologie; voir DEFRISE-GUSSENHOVEN, E., 1955.

Résumé.

Le problème d'analyse multivariée traité dans cette note s'est posé lors de la confrontation d'un fémur fossile avec une collection de fémurs actuels.

L'auteur définit une mesure de divergence entre un point fixe D (dont les coordonnées $d_1,...,d_p$ sont les mesures du fémur fossile) et un point M (dont les coordonnées $m_1,...,m_p$ sont les moyennes de la population de fémurs actuels).

Cette mesure de divergence est

$$\Lambda^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij}(m_i - d_i)(m_j - d_j) \qquad (i,j=1,...,p)$$

où les α^{ij} sont les mineurs normés des éléments α_{ij} de la matrice des covariances de la population que l'on suppose distribuée normalement.

La distance Λ^2 tient compte des différences $m_i - d_i$, mais aussi des covariances de la population. Dans le cas où $\alpha_{ij}=0$ $(i \neq j)$ et $\alpha_{ii}=1$, Λ^2 représente le carré d'une distance géométrique ordinaire.

Λ^2 ressemble à la distance généralisée que P. C. Mahalanobis utilise pour mesurer l'écart entre les moyennes de deux populations multivariées normales à matrices de covariances identiques.

Lorsque le point D varie, Λ^2 est distribué comme χ^2 avec p degrés de liberté. A chaque point D correspond une probabilité particulière qui indique le pourcentage de sujets de la population plus rapprochés de M que D. Pour des raisons d'ordre pratique, l'auteur appelle ce pourcentage « taux d'éloignement du point D ».

On peut utiliser Λ^2 pour comparer entre elles les distances de différents points D à une même population. On trouve un autre domaine d'application de Λ^2 dans le problème qui consiste à choisir entre plusieurs populations pour y ranger un spécimen, dont les mesures sont les coordonnées du point D : D appartiendra à la population à laquelle correspond le plus petit Λ^2. Ce procédé est intéressant parce qu'il ne nécessite pas l'identité des matrices de covariances des différentes populations.

L'auteur utilise certains théorèmes donnés par C. R. Rao pour établir la distribution échantillonnée de L^2 (valeur de Λ^2 calculée pour un échantillon).

Comme dans tous les cas apparentés, la distribution de L^2 ne dépend que de la valeur de Λ^2 pour la population, de l'effectif de l'échantillon et du nombre p de caractère envisagés.

Les premier et second moments de L^2 et un estimateur non biaisé et consistant de Λ^2 sont donnés. Pour calculer rapidement les intervalles de confiance de cet estimateur de Λ^2, il faudrait disposer d'une tabulation de la répartition pour différentes valeurs des paramètres.

A la connaissance de l'auteur, la distribution de L^2 est inédite. Elle est étroitement apparentée au type C donné par R. A. Fisher pour la distribution du coefficient de corrélation multiple *(loc. cit.)*.

La distribution de L^2 peut aussi être utilisée pour tester si D appartient à la population. La quantité $(n-1)L^2 = T^2$ pourrait être appelée T^2 non central de Hotelling; sa distribution serait celle de $(n-1)L^2$.

SUMMARY.

A problem of multivariate analysis has arisen in connection with a question of human paleontology in which a fossil bone was compared to a population of recent bones.

The author evaluates the divergence between a fixed point D (whose coordinates $d_1,..,d_p$ are the measures of a fossil) and a point M (whose coordinates $m_1,..,m_p$ are the means of the population of recent bones).

The defined measure of divergence is

$$\Lambda^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha^{ij}(m_i - d_i)(m_j - d_j) \qquad (i,j = 1,...,p)$$

where α^{ij} are the elements of the matrix reciprocal to the common dispersion matrix of the population, which is supposed to be normally distributed.

The measure Λ^2 takes into account the differences $m_i - d_i$, but also the covariances of the population. It represents the square of an usual geometrical distance when $\alpha_{ij}=0$ $(i \neq j)$ and $\alpha_{ii}=1$.

Λ^2 is akin to the generalized distance of MAHALANOBIS given as a measure of the distance between the means of two multivariate normal populations which have identical dispersion matrices.

Λ^2 being distributed as χ^2 with p degrees of freedom for varying D, a particular probability corresponds to each point D, indicating the percentage of the population nearer to the mean M than the point D. For practical reasons, the author has called this percentage « taux d'éloignement du point D ».

Λ^2 may be used to compare the distances from several points D to the same population or when one must decide to which of several normally distributed populations the measures of D belong. In the latter case, D will be enlisted in that population to which the smallest value of Λ^2 corresponds. The advantage of this discriminatory process is that the considered populations need not have the same dispersion matrices.

To establish the sampling distribution of the sample value L^2 of Λ^2, the author has used some theorems given by C. R. RAO.

As in all the analogous cases, the distribution of L^2 depends only on the population value of Λ^2, the number in the sample and the number p of measured characters.

The first and second moments of L^2 and an unbiased and consistent estimate of Λ^2 are given. The tabulation of the distribution ought to be made for practical use.

In other papers, the author has given numerical examples of the use of Λ^2.

To our knowledge, the sampling distribution of L^2 is new, but it is closely connected with the type C given by R. A. FIHSER for the distribution of the multiple correlation coefficient *(loc. cit.)*.

The distribution of L^2 may also be used to test if D belongs to the sampled population. The quantity $(n-1)L^2 = T^2$ might be called the non central T^2 of HOTELLING, and its distribution would be that of $(n-1)L^2$.

INDEX BIBLIOGRAPHIQUE.

BOSE, R. C. et ROY, S. N., 1938, *The distribution of the Studentized D^2-statistic.* (Sankhyã, Calcutta, vol. 4, part 1, pp. 19-37.)

CRAMÈR, H., 1946, *Mathematical methods of statistics.* (Princeton University Press, 1 vol., 574 p.)

DEFRISE-GUSSENHOVEN, E., 1952, *Discrimination de populations voisines. Etude biométrique.* (Bull. Inst. royal des Sc. natur. de Belgique, Bruxelles, tome XXVIII, n° 46, 34 p.)

DEFRISE-GUSSENHOVEN, E., 1955, *Ellipses équiprobables et taux d'éloignement.* (Bull. Inst. royal des Sc. natur. de Belgique, Bruxelles, tome 31, n° 26, 31 p.)

DEFRISE-GUSSENHOVEN, E., 1955, *Mesure de divergence et taux d'éloignement entre les moyennes d'une communauté de Carbonicola et les types du groupe Communis.* (Volume Jubilaire du Chanoine Demanet, Assoc. pour l'Etude de la Paléont. et de la Stratigr. Houillères, n° 21, Hors série, VIII, 418 p., 28 pl., Bruxelles.)

FISHER, R. A., 1915, *Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population.* (Biometrika, vol. X, pp. 507-522.)

FISHER, R. A., 1928, *The general Sampling Distribution of the Multiple Correlation coefficient.* (Proc. Roy. Soc. A, vol. 121, pp. 654-673.)

FISHER, R. A., 1936, *The use of Multiple Measurements in Taxonomic Problems.* (Annals of Eugenics, vol. 7 (2), pp. 179-188.)

FISHER, R. A., 1936, *Coefficient of Racial Likeness and the future of Craniometry.* (Journ. of the royal Anthr. Inst., Londres, vol. 66, pp. 57-64.)

FISHER, R. A., 1937, *The statistical utilization of multiple measurements.* (Annals of Eugenics, vol. 8, pp. 376-386.)

HOTELLING, H., 1931, *The generalization of « Student's » ratio.* (Ann. Math. Stats., vol. 2, pp. 360-378.)

MAHALANOBIS, P. C., 1930, *On tests and measures of group divergence.* (Journ. and Proc. Asiat. Soc. of Bengal, vol. 26, New Series, pp. 541-588, Calcutta.)

MAHALANOBIS, P. C., 1936, *On the generalized distance in Statistics.* (Proc. Nat. Inst. of Sc. of India, vol. 2 (1), pp. 49-55.)

RAO, C. R., 1952, *Advanced Statistical Methods in Biometric Research.* (John Wiley, New York, 1 vol., 389 p.)

ROY, S. N., 1939, *A note on the distribution of the Studentized D^2 Statistic.* (Sankhyã, Calcutta, vol. 4, part 3, pp. 373-380.)

SELTZER, C. C., 1937, *A critique of the coefficient of racial likeness.* (Amer. Journ. of Phys. Anthr., vol. 23, pp. 101-109.)

TWIESSELMANN, F., 1954, *Propos sur l'anthropologie.* (Volume jubilaire Victor Van Straelen, 1925-1954, tome II, pp. 1065-1098, Bruxelles.)

TABLE DES MATIÈRES.

INSTITUT ROYAL DES SCIENCES NATURELLES DE BELGIQUE.

AD. GOEMAERE, Imprimeur du Roi, 21, rue de la Limite, Bruxelles